JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MATSCREEN AI: A MULTI-SCALE GRAPH NEURAL NETWORK FRAMEWORK WITH EVIDENTIAL UNCERTAINTY QUANTIFICATION AND ENSEMBLE-MLIP CROSS-VALIDATION FOR RELIABILITY-FIRST MATERIALS DISCOVERY

[1]Dr. Vijayakumar K, [2]Matharishwa S, [3]Bichitra Behera, [4]Bharath S,

[5]Khushi Kalpesh Joshi

[1]Professor ,[2,3,4,5] Final Year Student

[1]Department of Computer Science Engineering (Data Science),

[1]AMC Engineering College, Bangalore, India

Abstract: **Pre-screening inorganic crystals with machine-learning models cuts the number of density functional theory (DFT) calculations a discovery campaign must pay for, but a graph neural network that emits a bare point prediction gives no warning when it is wrong. We describe MatScreen AI, a screening system that pairs a multi-scale property predictor with calibrated uncertainty and independent potentials check before anything reaches DFT. The predictor runs three message-passing branches at 4, 6, and 8 Å cutoffs, merges them with cross-attention, and emits a Normal-Inverse-Gamma posterior so that one forward pass yields both a formation-energy estimate and an interval. On a 50,000-structure Materials Project subset scored under a ten-fold leave-one-cluster-out SOAP protocol (48,998 held-out points) the model averages 0.0906 eV/atom mean absolute error (95% cluster-bootstrap CI 0.0796–0.1031), against 0.1962 eV/atom for a zero-shot ALIGNN reference on the same pool. Nested leave-one-fold-out conformal scaling holds 89.6% pooled coverage at a 0.1717 eV median half-width, and a distance-conditional quantile keeps the worst cluster at 78.2% coverage while staying within 1% of the interval width; it ships as an optional shift-aware mode. A dual-potential CHGNet/MACE gate on 177 ambiguous pairs places 98.3% of them in high-confidence agreement at a 3.2 meV/atom 90th-percentile disagreement. A Quantum ESPRESSO tier behind the gate provides PBE relaxations and gaps, a Δ-ML band-gap corrector that refuses out-of-domain queries, r2SCAN and HSE06 spot checks, and DFPT response; its ten regression tests reproduce pre-registered reference values in 259 s. A retrospective screening simulation on the held-out labels shows that ranking candidates by point-prediction proximity needs 6,424 expensive evaluations to recover 90% of true hits in the $|Ef| \in [1.2, 1.6]$ eV/atom window, versus 44,078 for random order, while a sigma-only promotion gate applied alone adds 18,565 evaluations. The simulation is a replay of stored labels, not a DFT result.**

IndexTerms **- Graph Neural Networks, Materials Discovery, Uncertainty Quantification, Evidential Deep Learning, Conformal Prediction, Machine-Learned Interatomic Potentials, Crystal Property Prediction, Density Functional Theory.**

I. INTRODUCTION

A discovery campaign starts with a chemical space far larger than any first-principles budget can cover. Full relaxations of complex unit cells run hours to days each, so every serious workflow puts a learned model in front of the simulator and lets it throw away most of the search space. Models trained on the Materials Project do this well enough to be standard equipment [1], [2], [9].

The failure mode is quiet. A network evaluated on a composition far from its training distribution still returns a formation energy, still returns it with a confident look, and the campaign then spends synthesis and simulation budget on phases that were never viable. In the worst case the pipeline reports a discovery that does not exist. Calibrated uncertainty changes that arithmetic: if the model reports how far it is reaching, routing (skip, hold, simulate) can act on the interval instead of a point estimate nobody can judge.

MatScreen AI builds three checks into the screening workflow:

1) Multi-scale receptive fields: Three parallel NNConv encoders read the same crystal graph through 4, 6, and 8 Å cutoffs, and a cross-attention head decides structure by structure how much weight the short, medium, and long range each deserve, instead of committing to a single radius for every crystal system [14].

2) Calibrated evidential intervals: The prediction head is trained with Deep Evidential Regression and returns a Normal-Inverse-Gamma posterior, so aleatoric and epistemic terms separate in one forward pass without ensemble or test-time sampling cost [4]. Split-conformal calibration then rescales those intervals against held-out scores, and we report both the deployed constant quantile and nested leave-one-fold-out quantiles that never see the fold they score [16].

3) Independent physical checks: Promoted candidates relax under two universal machine-learned interatomic potentials whose errors against DFT are largely uncorrelated, so disagreement between them is a model-independent reason to send a structure to DFT [6], [7], [8].

This paper reports the system after a full benchmarking round rather than a single headline number. Ten held-out chemical clusters score 48,998 points, each fold gets a quantile fitted on the other nine, and every number in the results section comes from committed artifacts that a compiler script re-derives. The contributions are: (i) the multi-scale evidential predictor and its ablation family; (ii) a nested conformal benchmark with a distance-conditional variant that improves the worst cluster at constant interval width; (iii) the dual-potential disagreement gate with an offset-free relaxation-drop metric; (iv) a retrospective screening simulation that prices routing policies in expensive evaluations needed for a target recall; and (v) the Tier-3 Quantum ESPRESSO validation stage, from PBE through HSE06, with a domain-gated Δ-ML band-gap corrector and a pre-registered regression suite.

II. RELATED WORK

Crystal property networks. CGCNN established the graph formulation for periodic crystals by turning pairwise distances into radial basis edge features [1]. ALIGNN added line graphs so three-body angles enter explicitly [2], and ComFormer replaced message passing with lattice-invariant attention over a single graph [3]. Our branch structure keeps the message-passing backbone and varies the cutoff instead, because bonding range is a property of the chemistry rather than of the architecture search.

Uncertainty for materials models. Deep ensembles and Monte Carlo dropout buy uncertainty at inference latency; deep evidential regression buys it at training time and returns a distribution from one pass [4]. The MatUQ benchmark introduced the SOAP leave-one-cluster-out protocol we adopt, showing that random splits report optimistic errors below 0.03 eV/atom while cluster holdouts are far harsher [5]. Conformal prediction adds a distribution-free coverage guarantee on top of any scoring model [16]; we use it in a nested form so the quantile for a fold is fitted only on the other folds.

Potentials as validators. Universal MLIPs relax structures quickly [6], [7], but single-potential screening inherits that potential's blind spots. Basu and Chakraborty report that relying on one potential misses most true DFT-stable crystals on comprehensive benchmarks, and that distinct architectures decorrelate their errors, which is what the disagreement gate uses [8].

First-principles tiers. Plane-wave DFT through Quantum ESPRESSO is what a screening pipeline eventually has to match [18], with PBE as the default functional [19] and screened-exchange HSE06 for band edges at several times the cost [20]; standardized pseudopotential sets such as SSSP remove one source of run-to-run drift [21]. The tier we describe is deliberately narrow: four stages, each regression-tested against pre-registered reference systems before it ever sees a candidate.

High-throughput and generative screening. GNoME and the A-Lab workflow showed the scale of autonomous screening and synthesis [9], [10]; retrospective audits of those campaigns argue for conservative triage [10]. MatterGen explores unconstrained structure generation [11], whereas our generator stays inside known scaffolds, and Matbench Discovery scores energy-above-hull predictions against a formal protocol [12]. We follow that benchmark's direction as future work rather than claiming stability from formation energy alone.

III. METHODOLOGY

A. Problem Formulation

Given a candidate crystal as a periodic graph G = (V, E) with species and fractional coordinates, the model must return a scalar formation energy per atom y (eV/atom) together with an interval that covers y at a stated rate, rather than a bare point estimate. Two further decisions ride on that output: whether the candidate differs structurally from anything already recorded in the reference database, and whether its prediction is trustworthy enough to act on without an independent calculation.

B. Multi-Scale Graph Neural Network Architecture

The backbone runs three NNConv-style encoder branches over the same graph with interaction radii 4, 6, and 8 Å. The tight radius resolves local coordination without pulling in distant shells; the wider radii carry lattice packing and longer-range packing effects that a 4 Å graph cannot see. A ScaleFusionAttention layer weights the three branch embeddings per structure before pooling, so the fusion is a learned function of the crystal rather than a fixed concatenation. Fig. 1(a) sketches the predictor.

C. Evidential Uncertainty Quantification

The fused embedding feeds a head that outputs the four parameters (γ, ν, α, β) of a Normal-Inverse-Gamma distribution [4]. The predictive mean is γ; the predictive variance splits into an aleatoric term $\beta/(\alpha - 1)$ and an epistemic term $\beta/(\nu - 1)$. Training minimizes the evidential negative log-likelihood plus a regularizer that damps overconfident evidence on high-residual samples. The delivered model is deterministic by default; a Monte Carlo dropout mode exists in the serving stack but is not used for the numbers reported here.

D. Conformal Calibration

Raw evidential intervals shift with training dynamics and dataset scale, so a held-out calibration score rescales them. With predictive standard deviation σ_i we use the normalized nonconformity score $s_i = |y_i - \hat{y}_i| / \sigma_i$ and a quantile $\hat{q}$ of those scores; intervals become $\hat{y}_i \pm \hat{q}\,\sigma_i$. Three quantiles are used:

• Shipped $\hat{q} = 1.0254$, fitted on the production calibration split and currently deployed.

• Nested LOFO, fitted on the other nine folds for each held-out fold. This is the primary reported configuration: it never scores a point with a quantile that point helped fit.

• Distance-conditional, a per-bin quantile. Each test point's distance to its own training set is computed as cosine distance between 512-dimensional count-sketch projections of SOAP descriptors; pooled deciles of that distance define bins, and each bin's quantile is a 90% score quantile on the other folds. A point far from anything seen in training therefore receives a wider interval without penalizing the well-covered majority.

The distance-conditional mode ships only if it raises both the minimum and the first-quartile per-fold coverage over the global nested quantile while keeping the median half-width within 3% of it; the evaluation in Section V meets that rule.

E. Discovery Pipeline

Fig. 1(b) shows the five stages. A user input constraint (a lead-free halide perovskite with formation energy below −1.5 eV/atom, say) parses into compositional limits and structural constraints. Candidates are built by substituting allowed elements into reference scaffolds and must pass a pymatgen charge-neutrality gate before inference. Tier-1 runs the multi-scale GNN and routes each candidate to Promote, Hold, or Excluded according to the calibrated interval around the energy threshold, which keeps borderline candidates from being discarded or advanced by accident. Before Tier-2, a Weisfeiler–Lehman graph hash checks novelty against the reference database; the measured collision rate falls from 0.58% at a 2.0 Å hash radius to 0.01% at 4.0 Å over 50,000 structures, and inspection of the residual pairs shows symmetry-equivalent entries rather than hash failures.

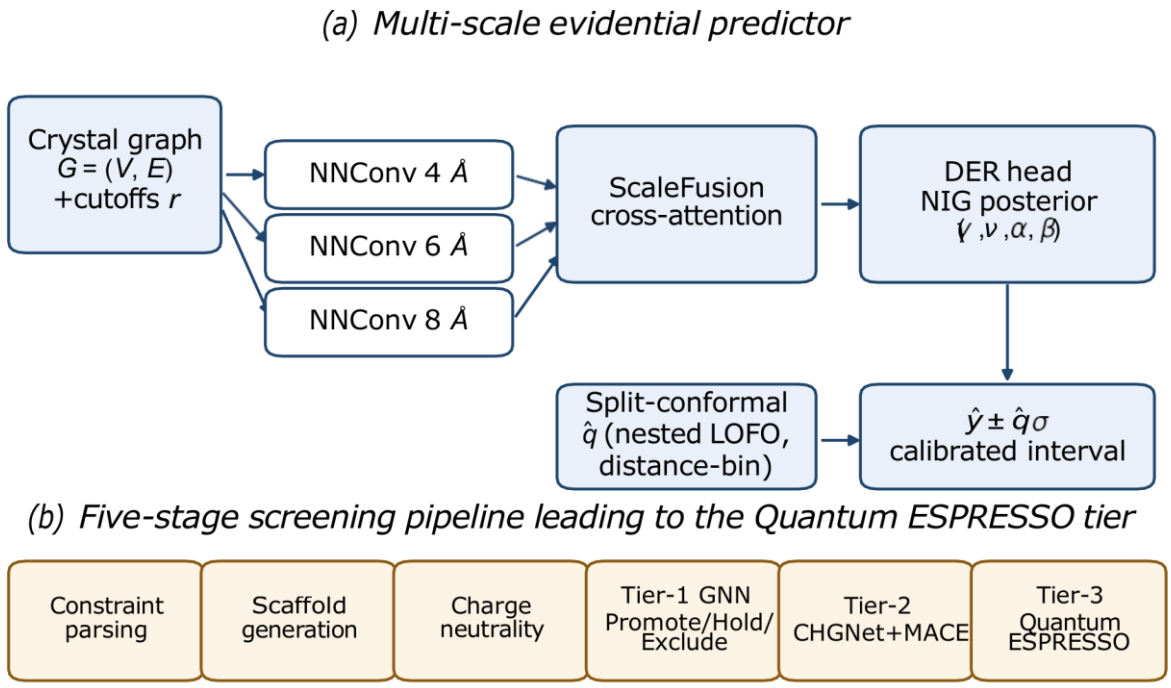

fig. 1. matscreen ai. (a) three nnconv branches at 4/6/8 å feed a cross-attention fusion whose embedding parameterizes a normal-inverse-gamma posterior; split-conformal scaling (nested lofo or distance-conditional quantiles) turns the posterior into a calibrated interval. (b) the five-stage screening pipeline; the ensemble-mlip gate is the last checkpoint before a candidate can be reported or handed to the tier-3 quantum espresso tier.

F. Ensemble-MLIP Disagreement Gate

Tier-2 relaxes each promoted candidate independently with CHGNet [6] and MACE [7] under ASE. The two potentials use different internal zero-point conventions, so raw total energies carry a 0.28–0.40 eV/atom offset that is not physical disagreement. We difference the per-model relaxation drop $\Delta E_i = E_{relaxed,i} - E_{initial,i}$, which cancels the offset, and compute the RMSD between the two relaxed geometries as a geometric check. The decision rule is: relaxation-drop difference at most 0.03 eV/atom with RMSD at most 0.10 Å counts as high-confidence agreement; a difference at or above 0.08 eV/atom, or RMSD at or above 0.10 Å, withholds the candidate for independent DFT validation.

G. Tier-2 Promotion Gate

Routing also needs a rule for which candidates deserve the expensive tier at all. The deployed orchestrator promotes a candidate when its predictive standard deviation reaches $\sigma \geq 0.09997$ eV, the threshold that marks a 50% promotion target on the pooled LOCO score distribution (it is also the pooled median sigma, so 24,499 of 48,998 points pass). The rule is stated in sigma units so it does not drift with the conformal quantile; the earlier interval-width rule, $\hat{q}\sigma \geq 0.10$ eV, would have promoted 99.7% of points at the shipped quantile and gave no selectivity at all.

H. Tier-3: Quantum ESPRESSO Validation

Anything the ensemble gate withholds for first-principles validation enters a Quantum ESPRESSO (QE) tier [18] with four ordered stages. Tier 3a runs a PBE [19] variable-cell relaxation, self-consistent field, and NSCF gap extraction using SSSP efficiency pseudopotentials [21] with per-element cutoffs; every Monkhorst–Pack subdivision is forced even and at least two, so zone-boundary points cannot be silently missed by an odd mesh, and high-symmetry cubic candidates get an explicit audit that the

relevant points are sampled. Tier 3b applies the Δ-ML band-gap corrector: a Ridge model on PBE gap plus composition and dielectric features, fitted on nine verified experimental optical gaps and evaluated leave-one-out. Tier 3c runs an r2SCAN single point on the PBE geometry for gaps that semi-local DFT visibly misplaces; Tier 3d evaluates HSE06 [20] eigenvalues at the VBM and CBM k-points only, which is where hybrid accuracy matters and nowhere else. A DFPT ph.x calculation supplies Born charges and the dielectric tensor for candidates that reach the phonon service.

The corrector applies domain gates before it answers. A PBE metal returns pbe_metallic_hse_undetermined instead of a number; a PBE+U system such as LiCoO2 is disclosed as outside the Δ-ML training domain; tin halides are rejected as a separate spin–orbit regime; and in-domain lead halides carry an 80% conformal interval ($\tilde{q}$ = 0.2157 eV) rather than 90%, because six calibration residuals cannot support a 90% order statistic ($k = \lceil 7 \times 0.9 \rceil = 7 > 6$). Open-shell transition-metal halides like KZrCl3 return pending_metallicity_check, so the candidate's gap field stays empty until the QE run resolves it.

IV. EXPERIMENTAL SETUP

Dataset and splits. Training and evaluation use a 50,000-structure subset of the Materials Project. Following the MatUQ protocol [5], structures cluster under SOAP descriptors and each fold holds out entire clusters, so scoring measures transfer to unseen chemistry instead of interpolation. The benchmark scores 48,998 held-out points across ten folds; fold sizes range from 101 to 7,678.

Metrics. Alongside MAE and RMSE we report Gaussian negative log-likelihood (NLL), the 15-bin normalized-uncertainty expected calibration error (ECE) used by the repository's metric suite, Spearman rank correlation between σ and $|y - \hat{y}|$, empirical coverage at each quantile, and selective-prediction risk, meaning MAE on the fraction of points retained when candidates are ranked by ascending sigma.

Baselines and ablations. A zero-shot ALIGNN evaluation exists on the same pool (13,555 scored rows, MAE 0.1962 eV/atom) and is reported as context only: it was not retrained under the same budget, so it is not a same-budget comparison. The MatUQ paper's ALIGNN figure of 0.046 eV/atom comes from a narrower perovskite subset [5] and is shown for context, not as a head-to-head number. No same-budget retrained CGCNN or ALIGNN baseline was completed in this window. Ablations A1–A7 cover single-scale, no-fusion, and no-evidential variants of the model on the same split at seed 42, one seed each.

Ensemble validation campaign. 177 pairs previously flagged as WL-hash collisions (290 unique structures) relax under both potentials; the whole campaign runs 505.6 s wall-clock.

Retrospective screening simulation. This experiment replays stored LOCO held-out labels; no DFT is run and no band gap is predicted. A hit is a label inside one of three windows fixed before the run: $|E_f| \in [1.2, 1.6]$ eV/atom, $|E_f| \in [2.0, 2.4]$ eV/atom, and the one-sided $E_f \leq -2.0$ eV/atom. Pooled hit counts are 4,378, 6,126, and 17,961. Four policies order candidates for expensive evaluation: random order; pure point-prediction proximity to the window; promoted-first ordering under the 50% sigma gate; and promoted-first ordering under an interval-overlap gate that promotes a candidate when $[\hat{y} \pm \hat{q}\sigma]$ intersects the window at the shipped, nested, or distance-conditional quantile. Both groups inside a gated policy are ordered by point-proximity, never at random. The hard-block operating point evaluates only promoted candidates and accepts non-promoted candidates without evaluation whenever the point prediction falls inside the window. Reported counts are the expensive evaluations needed to reach 90% and 95% recall of the true hits, per fold and pooled, with 95% percentile bootstrap intervals (B = 1000; points resampled within a fold, folds resampled as clusters for pooled numbers).

Quantum ESPRESSO regression suite. The QE tier is checked by ten tests marked qe against pre-registered reference values: Si (0.55–0.95 eV, indirect), MgO (4.35–4.95 eV, direct at Γ) plus its deformation-potential slope, NaCl (4.85–5.45 eV, direct), Cu (metallic), LiCoO2 with U_Co = 3.32 eV (must not collapse to a metal), and NaCl DFPT (ϵ_∞ in 2.30–2.70, Born charges within ±[1.00, 1.20] with the acoustic sum rule). The remaining three guard the Δ-ML corrector's leave-one-out accuracy and domain boundaries and the even-k-mesh guarantee. All ten pass; the suite runs in 259 s.

Held-out QE candidates. Two candidates run the full PBE pipeline at fixed literature lattices: FAPbI3 (a = 6.362 Å, target 1.48 eV) and MASnI3 (target 1.20 eV).

V. RESULTS

A. Property Prediction Accuracy

Table 1 gives the ten-fold result and Fig. 2 plots it. Mean MAE is 0.0906 eV/atom with a cluster-bootstrap interval of 0.0796–0.1031, median 0.0881, and the worst cluster (fold 7) reaches 0.1296. Mean RMSE is 0.1214 eV/atom (0.1051–0.1407). The spread between the best cluster (0.0608) and the worst (0.1296) is what chemical shift costs, and calibration below is reported per fold for that reason. Against the zero-shot ALIGNN reference at 0.1962 eV/atom, the fold-0 headline evaluation of the full configuration records 0.0641 eV/atom, a 3.06× reduction, with the zero-shot caveat above.

table 1. ten-fold leave-one-cluster-out accuracy (production scoring arrays)

Fold	n	MAE (eV/atom)	RMSE (eV/atom)	cov. (nested)
0	7,178	0.0741	0.0985	0.986
1	6,898	0.0789	0.1060	0.890
2	2,527	0.0855	0.1106	1.000
3	7,678	0.0908	0.1229	0.883
4	6,103	0.0907	0.1213	0.897
5	101	0.0958	0.1157	1.000
6	5,332	0.0608	0.0825	0.957
7	6,066	0.1296	0.1746	0.751

Fold	n	MAE (eV/atom)	RMSE (eV/atom)	cov. (nested)
8	6,604	0.0818	0.1080	0.855
9	511	0.1176	0.1740	1.000
mean/ pooled	**48,998**	**0.0906**	**0.1214**	**0.896**

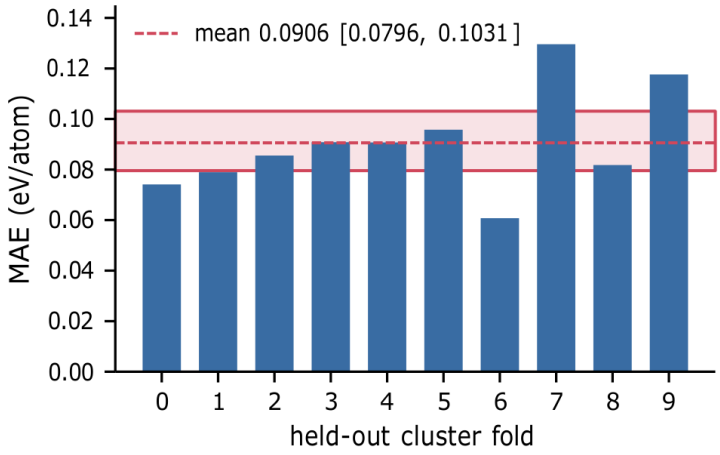

fig. 2. mae per held-out cluster fold. the dashed line and band mark the mean and its 95% cluster-bootstrap interval; fold 7 is the shifted cluster.

Table 2 lists the ablation family from the committed single-seed evaluation artifacts. The evidential, fused, multi-scale configuration (A7) is the one deployed; the no-DER variants trade away calibrated intervals for lower point error in these single-seed runs, and A5 has no committed evaluation and is marked as such. One seed per row means the table supports ordering observations, not significance claims.

table 2. ablation family on the soap-loco split, single seed 42

Config	MAE (eV/atom)	RMSE (eV/atom)	ECE	Spearman	cov.
A1 (4 Å single)	0.0369	0.0540	0.0323	0.199	0.916
A2 (6 Å single)	0.0364	0.0543	0.0534	0.192	0.957
A3 (8 Å single)	0.0393	0.0563	0.0233	0.209	0.958
A4 (Concat fusion)	0.0423	0.0580	0.0544	0.178	0.951
A5 (Sum fusion)	— (not evaluated in committed artifacts)	—	—	—	—
A6 (No DER, MSE)	0.0491	0.0683	0.0635	0.238	0.964
A7 (ScaleFusion DER)	0.0641	0.0831	0.0921	0.193	0.938

B. Calibration Under Distribution Shift

Fig. 3 and Table 3 compare the quantiles. The deployed constant $\hat{q}$ = 1.0254 covers 71.2% pooled and 78.4% macro at a 0.1025 eV median half-width; it is calibrated to the development split and under-covers the cluster-shifted benchmark, with fold 7 at 55.5%. The nested LOFO quantile (spread 1.6178–1.8165 across folds) restores 89.6% pooled and 92.2% macro coverage at 0.1717 eV half-width, which is the primary configuration. The distance-conditional variant gives up 0.6 points of pooled coverage but lifts the minimum fold from 75.1% to 78.2% and the first quartile from 0.8845 to 0.8911 at a 0.9937 width ratio, satisfying the shipping rule as an optional shift-aware mode. The own-fold shift-aware quantile (0.9123 pooled) is a non-nested oracle and is not claimed as a result. On an external lead-halide set of 235 structures the shipped quantile covers 92.8%, but the median sigma there is 2.51× the LOCO value, so that coverage comes from inflated uncertainty rather than transferred calibration. The LOCO benchmark's own 122 lead-halide points cover 60.7% at the same quantile.

table 3. coverage and interval width by quantile choice

Quantile Configuration	Pooled Coverage	Macro Coverage	Median Half-Width (eV)
shipped (q = 1.0254)	0.712	0.784	0.1025
nested LOFO (primary)	0.896	0.922	0.1717
pooled 10-fold (q = 1.7242)	—	0.925	0.1724
distance-bin (nested)	0.890	0.917	0.1707

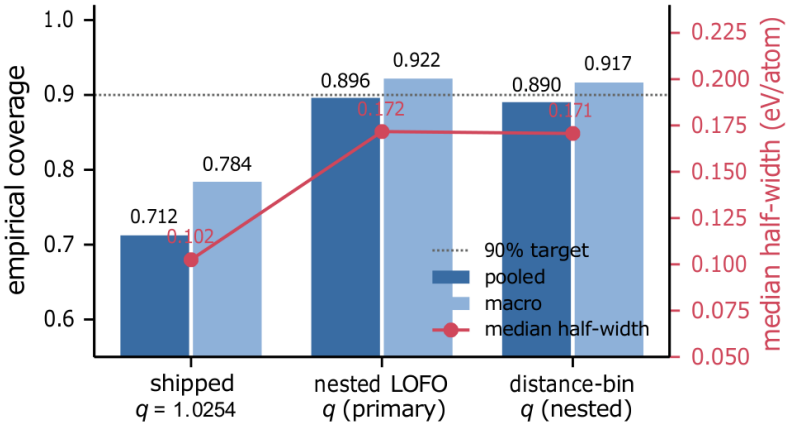

fig. 3. empirical coverage (bars, left axis) and median interval half-width (line, right axis) for the shipped, nested, and distance-conditional quantiles.

C. Predictive Quality and Selective Prediction

Computed directly from the committed score arrays, the pooled Gaussian NLL is −0.6621, pooled ECE 0.0578, and pooled Spearman (σ, |err|) 0.2571. Per fold, ECE ranges from 0.0728 (fold 6) to 0.7699 (fold 9, 511 points), and Spearman ranges from −0.1755 (fold 2) to 0.3956 (fold 7), so sigma ranks error acceptably almost everywhere but badly on the two folds where it should be treated as uninformative. All four metrics are computed from the stored target, mu, and sigma arrays.

The selective curve in Fig. 4 shows what the ranking is worth operationally. Keeping only the 10% of pooled candidates with the smallest sigma cuts MAE to 0.0523 eV/atom against 0.0870 at full retention, and the curve rises monotonically, so discarding the high-sigma tail buys a predictable accuracy gain on every fold. Fold 2 is the one-fold whose curve runs the other way, which matches its negative Spearman value.

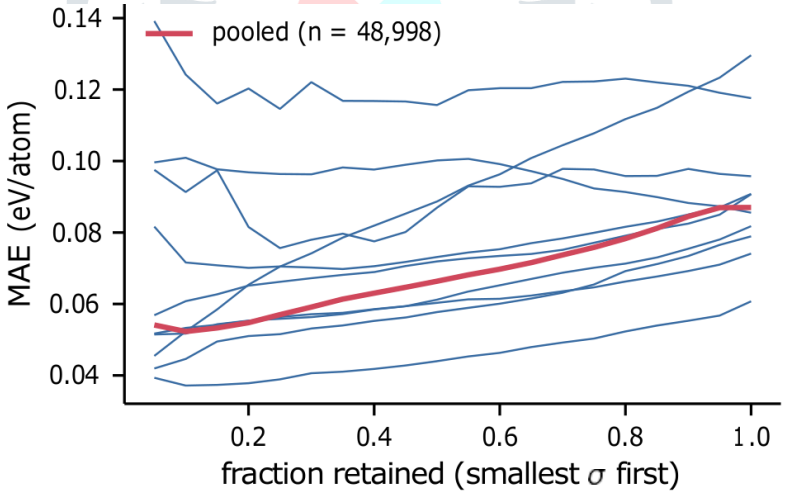

fig. 4. selective-prediction curve: mae against the fraction of candidates retained when ranked by ascending sigma. thin lines are individual folds; the thick line is the pooled pool of 48,998 points.

D. Ensemble-MLIP Validation Campaign

Across the 177 collision pairs, 174 (98.3%) land in high-confidence agreement and 3 (1.7%) in moderate agreement; none trip the independent-validation flag, which is expected because the set already survived an earlier CHGNet screening pass. Relaxation-drop disagreement in meV/atom has mean 1.5, median 0.5, 90th percentile 3.2, and maximum 26.2. Geometry agrees to a 0.0063 Å median RMSD (90th percentile 0.0299 Å, maximum 0.1294 Å).

The gate also has to reject candidates. Table 4 shows four newly generated compositions run through the same thresholds: KZrCl3 and LiCrO2 settle into high-confidence agreement, KMnF2 lands in the moderate band, and MgMoF2 shows a 0.1360 eV/atom relaxation-drop difference, roughly five times the campaign's 90th percentile, which sends it to DFT rather than into the report.

table 4. discriminating-power test on generated candidates (relaxation-drop difference and rmsd)

Candidate	ΔE (eV/atom)	RMSD (Å)	Outcome Decision
KZrCl3	0.0000	0.0000	High confidence (verified)
LiCrO2	0.0158	0.0022	High confidence (verified)
KMnF2	0.0504	0.0070	Moderate (flagged review)
MgMoF2	0.1360	0.0026	Requires DFT validation

E. Retrospective Screening Simulation

The numbers below are a retrospective simulation on held-out formation-energy labels, not a DFT result and not a band-gap experiment.

Table 5 gives the pooled expensive-evaluation counts. Point-proximity ranking is close to free relative to random: 6,424 evaluations instead of 44,078 to reach 90% recall in the |Ef| ∈ [1.2, 1.6] window, and 16,288 instead of 44,096 in the one-sided window. The sigma-only Tier-2 gate, applied as a promoted-first ordering, is worse than pure proximity everywhere (+18,565 evaluations at 90% recall in the first window, cluster-bootstrap interval +8,161 to +27,104; the other windows agree). The gate promotes exactly the points whose sigma sits above the pooled median, so confident hits wait at the back of the queue. The interval-overlap gates tie with pure proximity at 90% recall (deltas within −52 to +7 evaluations, intervals overlapping zero) and stay statistically indistinguishable at 95%, with the shipped-quantile gate 2,163 evaluations ahead in the second window but an interval that still crosses zero. No gate was tuned after the fact.

table 5. retrospective simulation, pooled counts to recall (estimate [95% cluster-bootstrap ci])

Policy	W1 90%	W1 95%	W2 90%	W2 95%	W3 90%	W3 95%
random	44,078 [27.5k, 57.0k]	46,533 [29.0k, 60.2k]	44,108 [27.5k, 57.1k]	46,534 [28.9k, 60.2k]	44,096 [27.6k, 57.2k]	46,542 [29.0k, 60.2k]
point	6,424 [3.5k, 9.8k]	7,786 [4.3k, 12.0k]	7,740 [4.1k, 11.6k]	9,268 [5.1k, 13.7k]	16,288 [7.3k, 26.1k]	17,414 [8.0k, 27.7k]
σ-gate	24,989 [11.9k, 34.8k]	25,333 [14.7k, 35.5k]	26,842 [11.1k, 36.1k]	27,482 [17.4k, 37.1k]	31,013 [19.3k, 41.2k]	31,918 [20.1k, 42.5k]
int-shipped	6,414 [3.5k, 9.8k]	9,949 [4.5k, 14.2k]	7,740 [4.1k, 11.6k]	9,062 [5.2k, 13.6k]	16,288 [7.3k, 26.1k]	17,414 [8.0k, 27.7k]
int-nested	6,424 [3.5k, 9.8k]	7,764 [4.3k, 11.8k]	7,740 [4.1k, 11.6k]	9,262 [5.0k, 13.7k]	16,288 [7.3k, 26.1k]	17,414 [8.0k, 27.7k]
int-dist	6,422 [3.5k, 9.8k]	7,750 [4.3k, 11.8k]	7,740 [4.1k, 11.6k]	9,203 [4.9k, 13.7k]	16,288 [7.3k, 26.1k]	17,414 [8.0k, 27.7k]

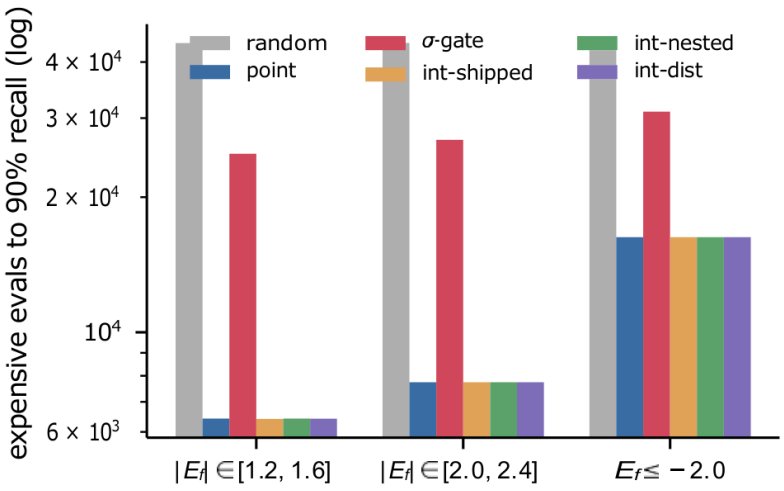

fig. 5. pooled expensive evaluations to 90% recall per policy and window (log scale). point-proximity and the interval gates sit far below random; the sigma-only gate sits far above them.

The hard-block operating point trades recall for skipped evaluations. In W1 the sigma gate declares 24,499 evaluations, reaches 94.3% recall at 96.8% precision, and cannot reach a 95% target at all; the shipped-quantile gate declares 9,588 evaluations at 100.0% precision but caps at 93.9% recall; the nested and distance-conditional gates declare 11,490 and 11,540 evaluations, both above 98% recall, and are the only W1 gates that clear 95% (7,764 and 7,750 evaluations inside the promoted set). In W3 every gate exceeds 99% recall, because a one-sided window at the far tail of the distribution is easy to cover. Per-fold tables, including folds where a window contains no hits, are in the committed benchmark report.

F. Quantum ESPRESSO Tier and Δ-ML Corrector

All ten QE-marked regression tests pass in 259 s: the PBE gaps for Si, MgO, and NaCl land inside their pre-registered windows (MgO and NaCl direct at Γ), Cu is correctly returned as metallic with no gap value, LiCoO2 keeps a gap once U_Co = 3.32 eV is applied instead of collapsing to a metal, and the NaCl DFPT run reproduces ϵ_∞ and both Born charges with the acoustic sum rule satisfied. The MgO deformation-potential check confirms the gap closes under lattice expansion with a slope in the literature window, which is the guard against a bad prototype lattice parameter propagating silently into every ionic insulator downstream.

On the held-out candidates, QE-direct PBE gives FAPbI3 a 1.2656 eV gap at the literature lattice (a = 6.362 Å, 104.4 s wall-clock); the frozen Δ-ML corrector maps it to 1.5648 eV against the 1.48 eV target, an error of 0.0848 eV, inside its 80% interval [1.2544, 1.8751] eV. The same corrector refuses MASnI3: its PBE gap of 0.3482 eV falls in the tin spin–orbit regime, and the domain gate answers out_of_domain with no number attached. Evaluated anyway, the ungated extrapolation misses by 0.7344 eV.

On the pooled nine-point fit with dielectric features, in-family halide LOOCV MAE is 0.2683 eV for the Ridge corrector against 0.6245 eV for a linear PBE scissor. The deployed halide-only fit (n = 6) is tighter still: anion-matched mean Δ reaches 0.1296 eV LOOCV against 0.3077 eV for a constant scissor and 0.1669 eV for the linear PBE scissor. These are small numbers on small samples, and Section VI says where they stop holding.

G. Case Study: KZrCl3

The pipeline's highest-ranked candidate is KZrCl3, an earth-abundant, lead-free halide perovskite with a predicted formation energy of −2.088 eV/atom and an estimated raw material cost of $13.63/kg. It passes the charge gate and shows exact high-confidence agreement between CHGNet and MACE at Tier-2. Secondary electronic properties for the candidate (band gap, dielectric response, carrier mobility) are labeled "requires DFT" in system output instead of being filled in by an unvalidated surrogate; an earlier prototype that emitted fixed default band gaps was removed for exactly that reason. For this candidate the Δ-ML corrector returns pending_metallicity_check (open-shell Zr in a halide perovskite is outside its calibration), so the gap field stays empty until the Tier-3 PBE run resolves it, with HSE06 available at the band edges if the PBE gap lands near a decision threshold.

VI. DISCUSSION AND LIMITATIONS

What the numbers do not say. Formation energy per atom is not synthesizability. Stable-phase screening needs energy above the convex hull against a reference phase diagram, and aligning with the Matbench Discovery protocol [12] is the immediate next step. The screening simulation in Section V-E replays historical labels: it prices routing policies, but no DFT was executed inside it, and its windows are formation-energy windows, not band gaps.

Benchmark gaps we did not close. The same-budget retrained CGCNN and ALIGNN baselines were not run in this window, so the zero-shot ALIGNN number stays labeled as such. The ablation table is single-seed, the DER lambda sweep only trained $\lambda = 0.001$ (no evaluation committed, no results claimed for the other values), and per-fold MC-dropout uncertainty was not committed.

Shift is real and only partly handled. Fold 7's cluster sits far enough from its training data that the shipped quantile covers 55.5% of it; nested and distance-conditional quantiles recover most, not all, of that gap (75.1% and 78.2%). Deployment keeps the shift-aware mode available and does not treat the problem as closed.

Operational discipline. The offset between CHGNet and MACE zero-point conventions and the removal of default band-gap values are both cases where a convenient number would have been a wrong number. The pipeline reports the disagreement or the missing quantity instead. Generation stays scaffold-constrained, so Stable-Unique-Novel rates computed from it are not comparable with unconstrained generators such as MatterGen [11]. The multi-scale cutoffs remain an empirical choice that worked in these runs, not a proven optimum, and the training loop does not yet re-ingest validated DFT outcomes automatically.

What the QE tier does not settle. The Δ-ML corrector is fitted on nine compounds, and one of them (MgO) is an extreme leverage point ($h = 0.9368$, Cook's $D = 1176$) whose removal changes the pooled fit; that is why non-halide queries return out_of_domain instead of an extrapolation. The 80% interval on in-domain lead halides is a floor set by six residuals, not a choice. Self-consistent PBE itself disagrees with published calculations on the same structures: our CsPbI3 gap is 1.323 eV against 1.14 eV from Wiktor et al., and the MAPbX3 series runs 0.19–0.24 eV below Mosconi et al. with grid density ruled out as the cause and the discrepancy still unisolated. Finally, the Δ bundles more than exchange–correlation error: at 300 K it also carries zero-point renormalization, thermal expansion, and dynamic disorder, so a corrected gap is an optical-gap estimate at experimental temperature, not a 0 K Kohn–Sham quantity.

VII. CONCLUSION AND FUTURE WORK

MatScreen AI couples a multi-scale evidential GNN with nested conformal calibration and a dual-potential gate, and reports what each piece is worth. The predictor averages 0.0906 eV/atom MAE across ten held-out chemical clusters (95% CI 0.0796–0.1031) with 89.6% pooled coverage at 0.1717 eV median half-width under the nested quantile; the distance-conditional variant holds that coverage while lifting the worst cluster to 78.2% at unchanged width. The CHGNet/MACE gate places 98.3% of 177 ambiguous pairs in high-confidence agreement and isolates MgMoF2 for DFT at 136 meV/atom disagreement. The retrospective simulation shows what routing is worth in evaluations: point proximity recovers 90% of hits in 6,424 evaluations where random order needs 44,078, and a sigma-only promotion rule costs 18,565 more evaluations on top of proximity, which is a reason to keep the deployed gate for uncertainty triage and not for hit ranking.

The near-term roadmap is convex-hull stability scoring toward Matbench Discovery, running the Tier-3 PBE pipeline on KZrCl3 to fill the gap field the corrector refuses to guess, an active-learning loop that feeds validated DFT outcomes back into training, and evaluation of the scaffold-constrained generator under the Stable-Unique-Novel metric with full reporting.

ACKNOWLEDGMENT

The authors express their gratitude to the Department of Computer Science and Engineering (Data Science) at AMC Engineering College, Bengaluru, for providing the computational infrastructure, laboratory resources, and GPU compute clusters necessary to execute the materials discovery benchmark experiments.

REFERENCES

[1] T. Xie and J. C. Grossman, 'Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties,' Phys. Rev. Lett., vol. 120, no. 14, p. 145301, 2018.

[2] K. Choudhary and B. DeCost, 'Atomistic line graph neural network for improved materials property predictions,' npj Comput. Mater., vol. 7, no. 1, p. 185, 2021.

[3] S. Yan et al., 'ComFormer: Complete and efficient graph transformers for crystal material property prediction,' in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[4] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, 'Deep evidential regression,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 14927-14937, 2020.

[5] L. Tan, P. Chen, M. Liu, X. Wang, J. Cen, and Q. Zou, 'Benchmarking GNNs for OOD materials property prediction with uncertainty quantification,' arXiv:2511.11697, 2025.

[6] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, 'CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling,' Nat. Mach. Intell., vol. 5, pp. 1031–1041, 2023.

[7] I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, and G. Csányi, 'MACE: Higher order equivariant message passing neural networks for fast and accurate force fields,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 11423–11436, 2022.

[8] A. Basu and P. Chakraborty, 'Proof-carrying materials: Falsifiable safety certificates for machine-learned interatomic potentials,' arXiv:2603.12183, 2026.

[9] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, 'Scaling deep learning for materials discovery,' Nature, vol. 624, pp. 80–85, 2023.

[10] N. J. Szymanski et al., 'An autonomous laboratory for the accelerated synthesis of novel materials,' Nature, vol. 624, pp. 86–91, 2023.

[11] C. Zeni et al., 'A generative model for inorganic materials design,' Nature, vol. 639, pp. 624–632, 2025.

[12] J. Riebesell et al., 'A framework to evaluate machine learning crystal stability predictions,' Nat. Mach. Intell., vol. 7, pp. 836–847, 2025.

[13] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, 'Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm,' npj Comput. Mater., vol. 6, p. 138, 2020.

[14] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, 'Neural message passing for quantum chemistry,' in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 1263-1272.

[15] S. P. Ong et al., 'A universal graph deep learning interatomic potential for the periodic table,' Nat. Comput. Sci., vol. 2, pp. 718–728, 2022.

[16] G. Shafer and V. Vovk, 'A tutorial on conformal prediction,' J. Mach. Learn. Res., vol. 9, pp. 371–421, 2008.

[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, 'A fast and elitist multiobjective genetic algorithm: NSGA-II,' IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002.

[18] P. Giannozzi et al., 'Quantum ESPRESSO: a modular and open-source software package for quantum simulations of materials,' J. Phys. Condens. Matter, vol. 21, no. 39, p. 395502, 2009.

[19] J. P. Perdew, K. Burke, and M. Ernzerhof, 'Generalized gradient approximation made simple,' Phys. Rev. Lett., vol. 77, no. 18, p. 3865, 1996.

[20] J. Heyd, G. E. Scuseria, and M. Ernzerhof, 'Hybrid functionals based on a screened Coulomb potential,' J. Chem. Phys., vol. 118, no. 18, p. 8207, 2003.

[21] G. Prandini et al., 'Precision and efficiency in solid-state pseudopotential calculations,' npj Comput. Mater., vol. 4, p. 72, 2018.